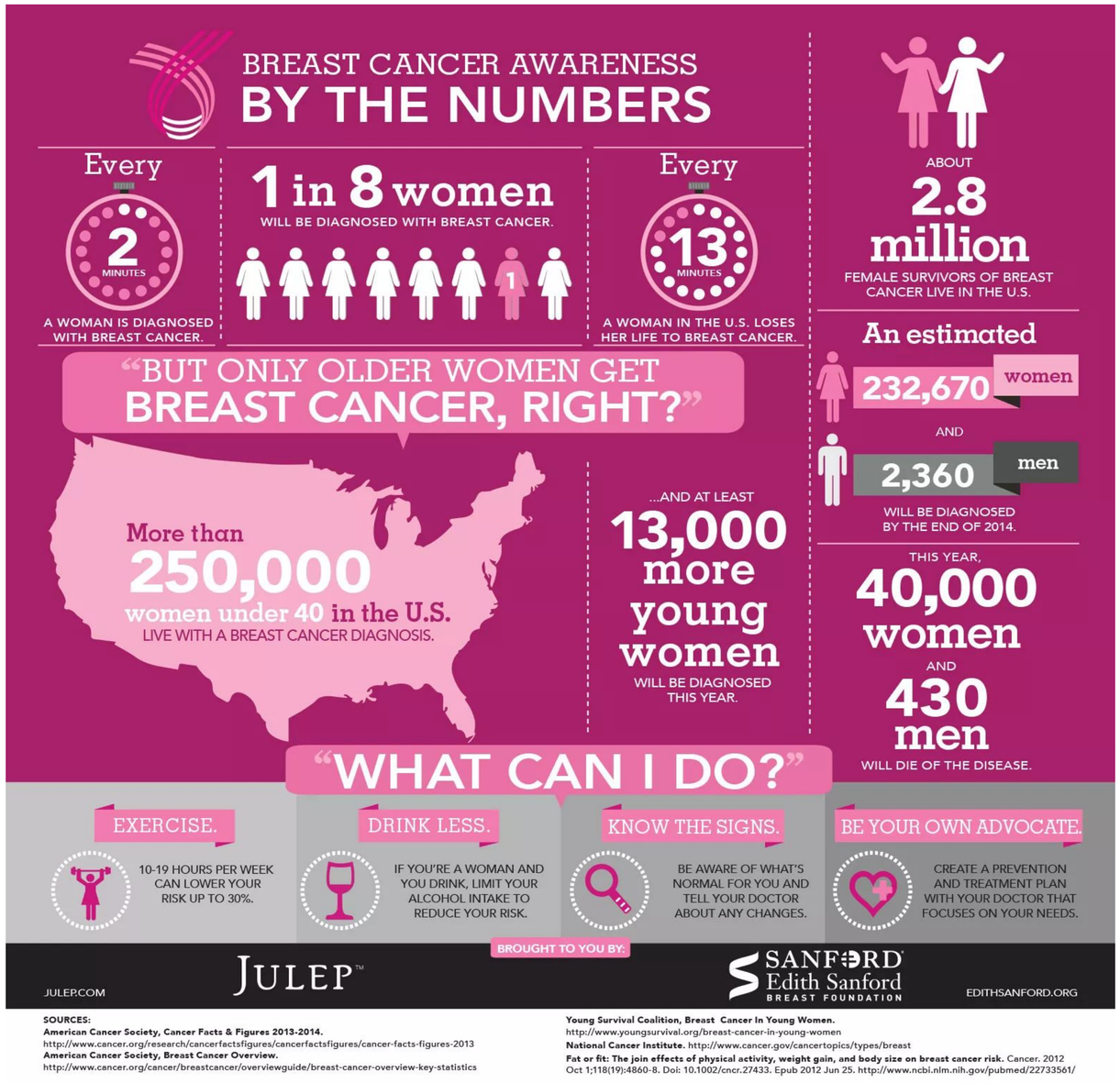
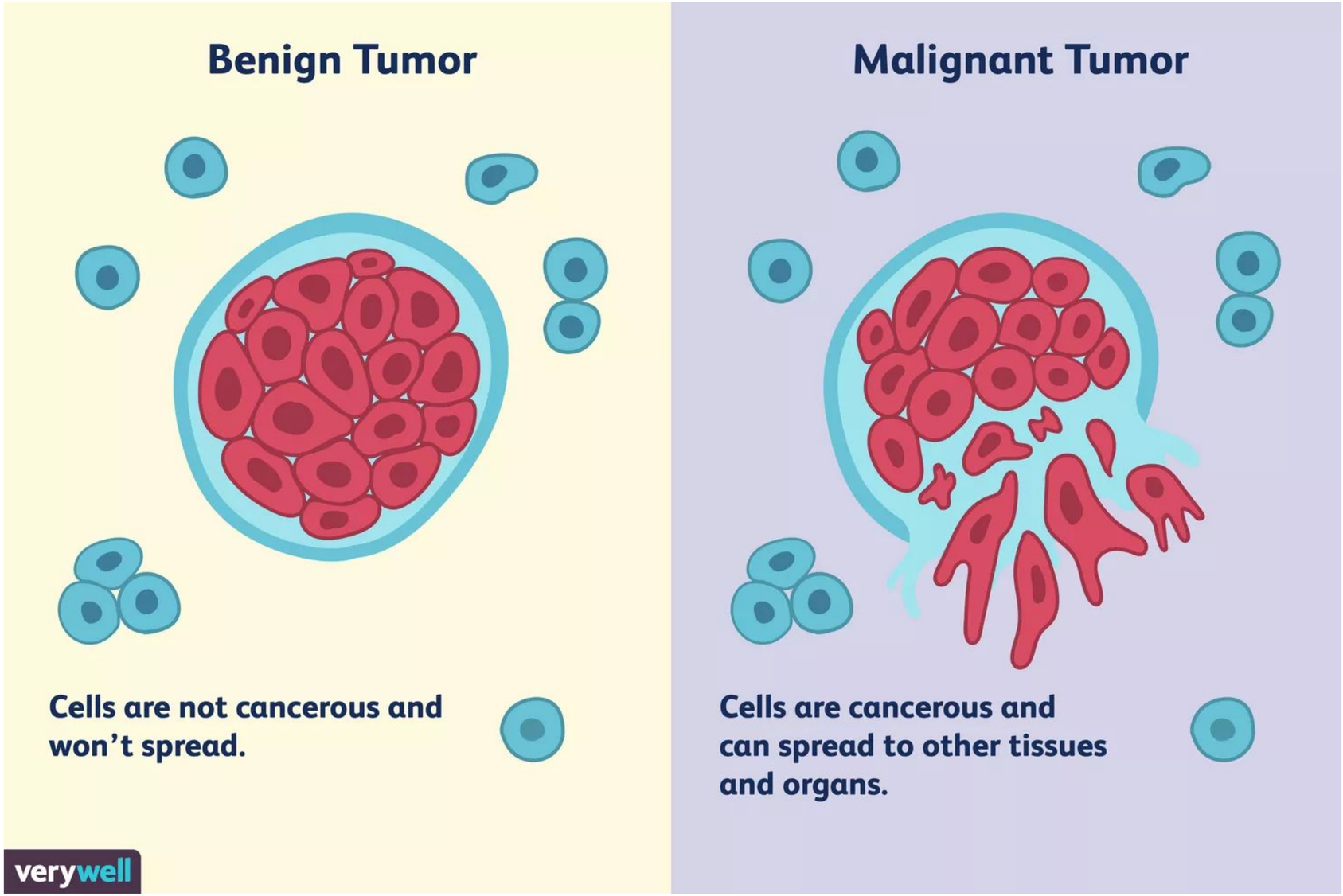
CS 106S: Week 4, Fall 2024

Breast Cancer Classification with k-Nearest-Neighbors

The Problem


BREAST CANCER AWARENESS
BY THE NUMBERS
Every
2
MINUTES
A WOMAN IS DIAGNOSED WITH BREAST CANCER.
1 in 8 women
WILL BE DIAGNOSED WITH BREAST CANCER.
Every
13
MINUTES
A WOMAN IN THE U.S. LOSES HER LIFE TO BREAST CANCER.
ABOUT
2.8 million
FEMALE SURVIVORS OF BREAST CANCER LIVE IN THE U.S.
An estimated
232,670 women
AND
2,360 men
WILL BE DIAGNOSED BY THE END OF 2014.
"BUT ONLY OLDER WOMEN GET BREAST CANCER, RIGHT?"
More than
250,000
women under 40 in the U.S.
LIVE WITH A BREAST CANCER DIAGNOSIS.
...AND AT LEAST
13,000 more young women
WILL BE DIAGNOSED THIS YEAR.
THIS YEAR,
40,000 women
AND
430 men
WILL DIE OF THE DISEASE.
"WHAT CAN I DO?"
EXERCISE.
10-19 HOURS PER WEEK CAN LOWER YOUR RISK UP TO 30%.
DRINK LESS.
IF YOU'RE A WOMAN AND YOU DRINK, LIMIT YOUR ALCOHOL INTAKE TO REDUCE YOUR RISK.
KNOW THE SIGNS.
BE AWARE OF WHAT'S NORMAL FOR YOU AND TELL YOUR DOCTOR ABOUT ANY CHANGES.
BE YOUR OWN ADVOCATE.
CREATE A PREVENTION AND TREATMENT PLAN WITH YOUR DOCTOR THAT FOCUSES ON YOUR NEEDS.
BROUGHT TO YOU BY:
JULEP™
JULEP.COM
SANFORD Edith Sanford BREAST FOUNDATION
EDITHSANFORD.ORG
SOURCES:
American Cancer Society, Cancer Facts & Figures 2013-2014.
http://www.cancer.org/research/cancerfactsfigures/cancerfactsfigures/cancer-facts-figures-2013
American Cancer Society, Breast Cancer Overview.
http://www.cancer.org/cancer/breastcancer/overviewguide/breast-cancer-overview-key-statistics
Young Survival Coalition, Breast Cancer In Young Women.
http://www.youngsurvival.org/breast-cancer-in-young-women
National Cancer Institute. http://www.cancer.gov/cancertopics/types/breast
Fat or fit: The join effects of physical activity, weight gain, and body size on breast cancer risk. Cancer. 2012 Oct 1;118(19):4860-8. Doi: 10.1002/cncr.27433. Epub 2012 Jun 25. http://www.ncbi.nlm.nih.gov/pubmed/22733561/

The Problem


Benign Tumor
Cells are not cancerous and won't spread.
Malignant Tumor
Cells are cancerous and can spread to other tissues and organs.
verywell

Today's Task

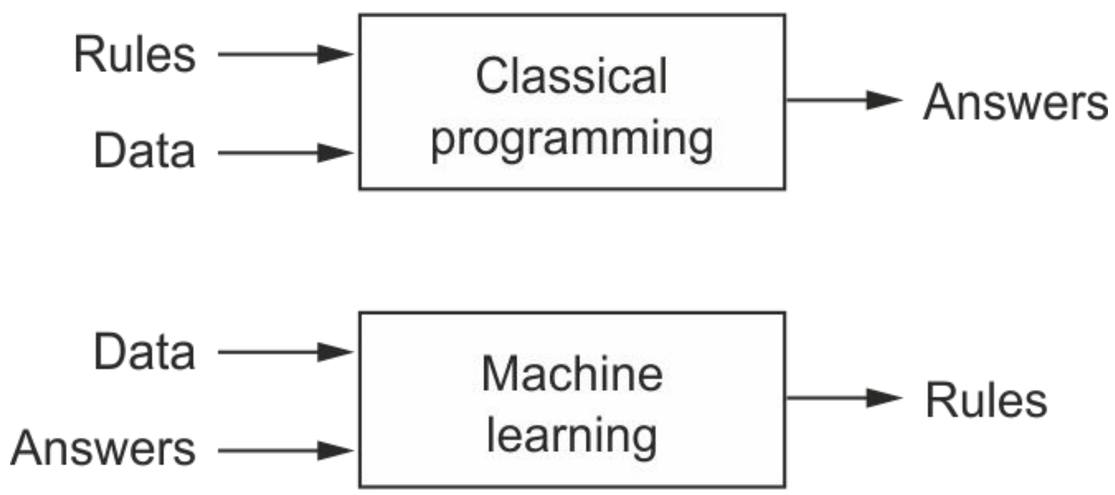
Given medical data about cell growths, can we accurately classify these tumors as benign or malignant?

First: A Review of Machine Learning

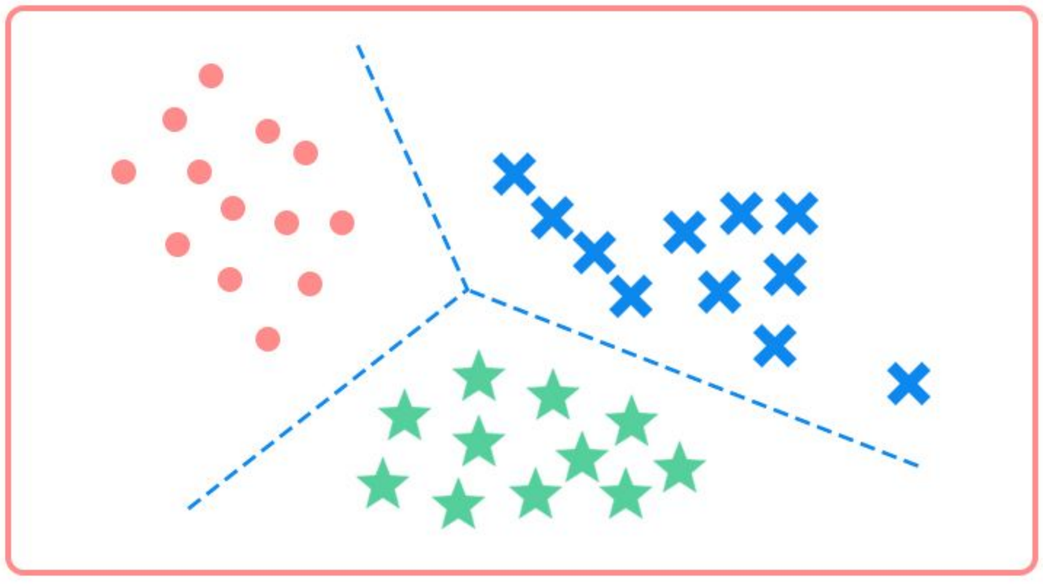
What is machine learning?

Machine learning is a “field of study that gives computers the ability to learn without being explicitly programmed.”

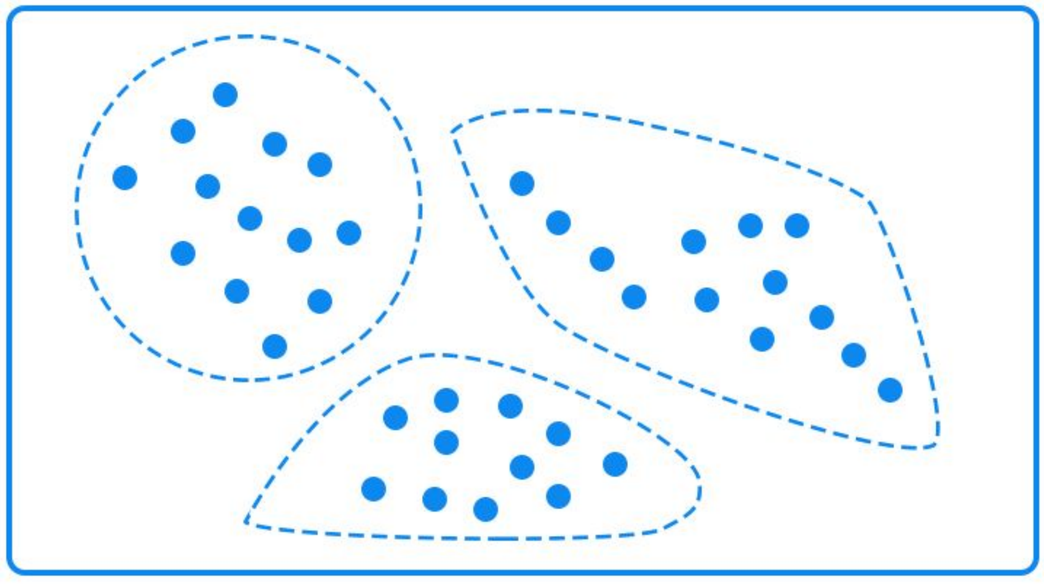
- Arthur Samuel, 1959



Supervised vs. unsupervised learning

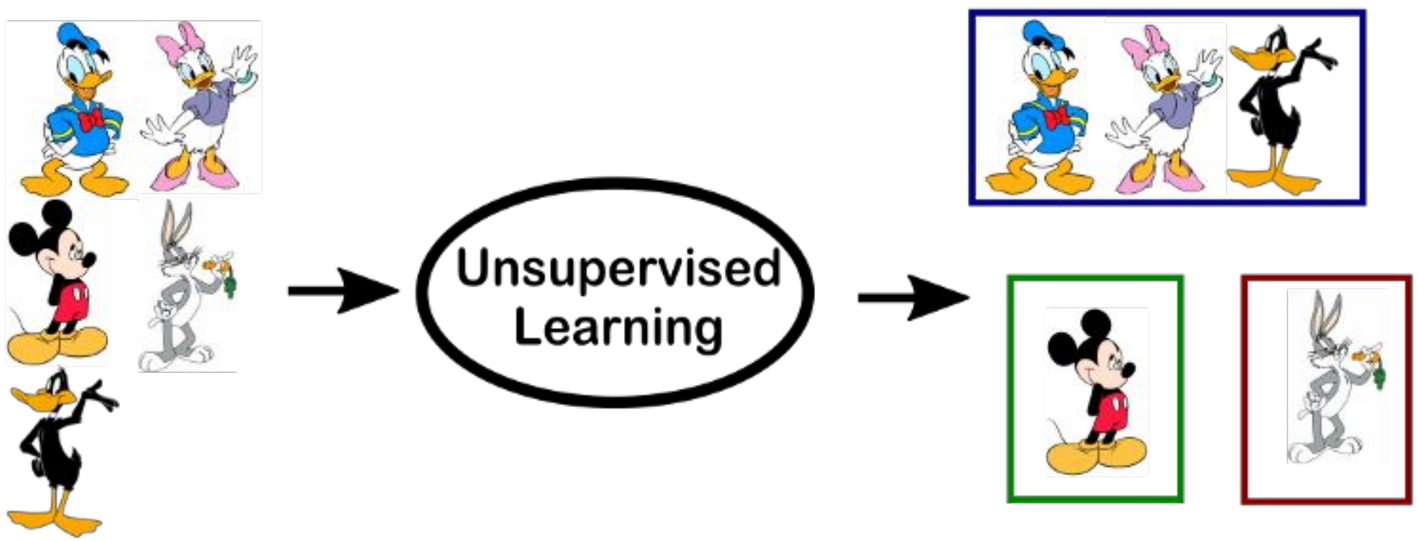


Supervised learning



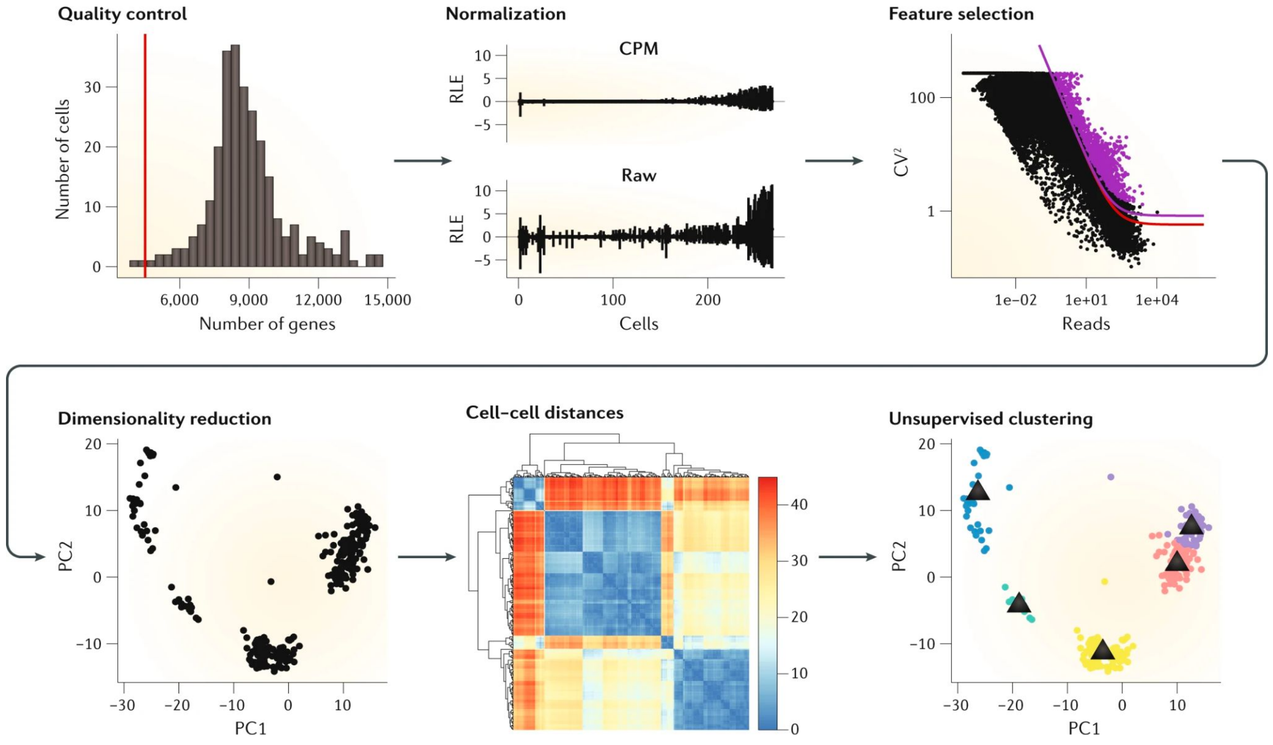
Unsupervised learning

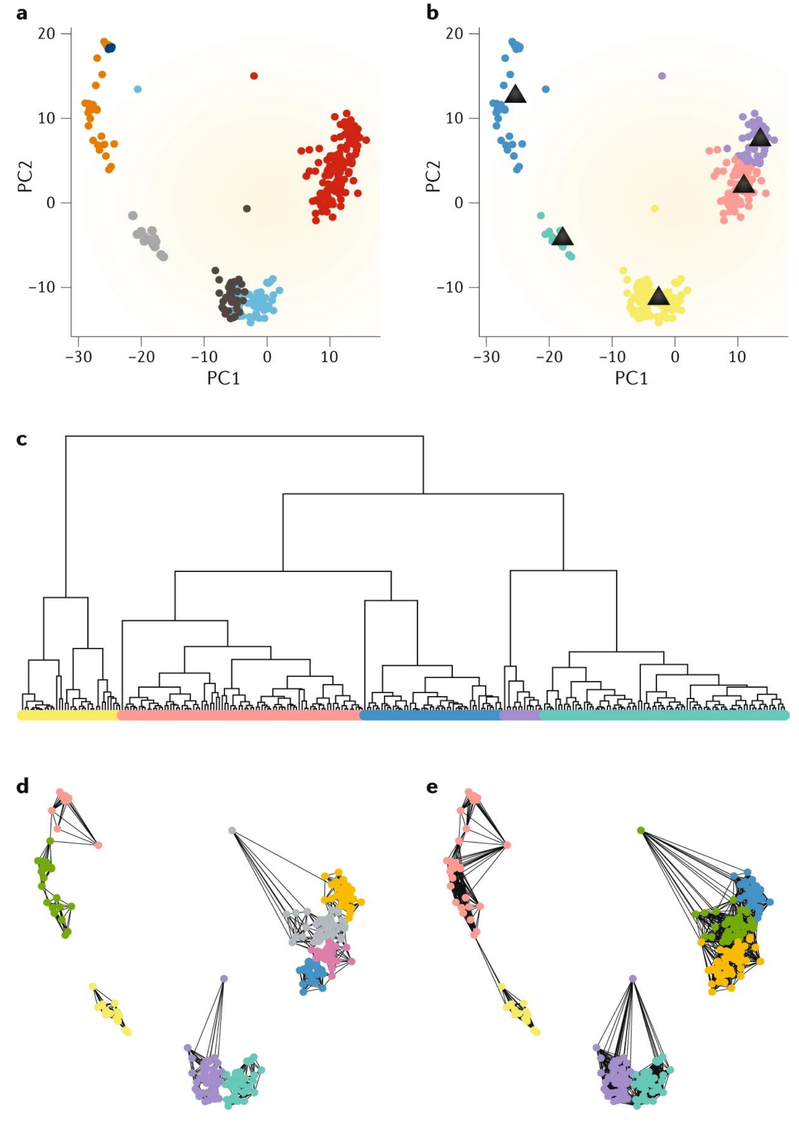
Unsupervised



https://www.nature.com/articles/s41576-018-0088-9/figures/3

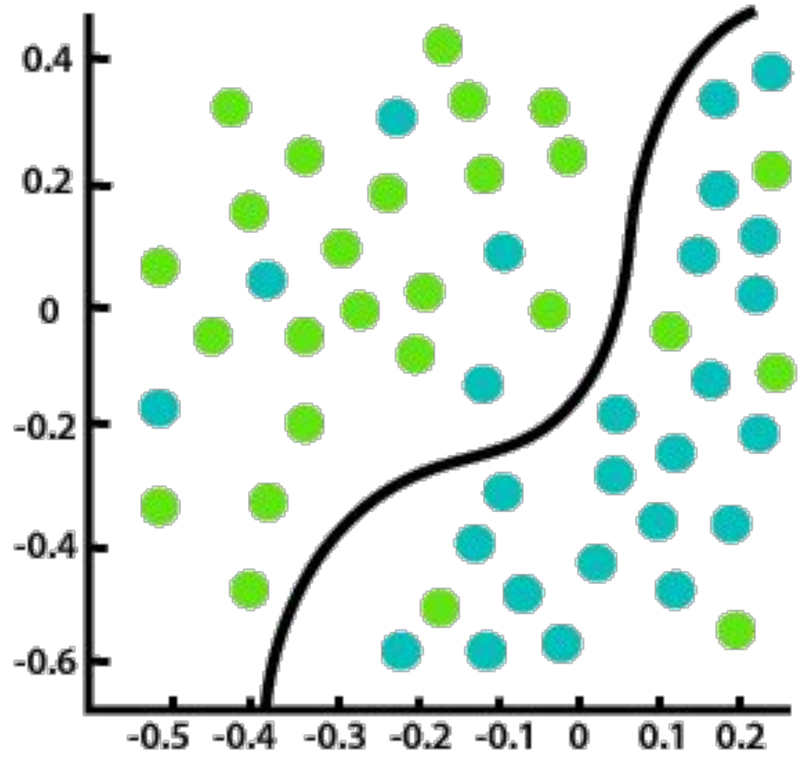
Example application: clustering scRNA-seq data



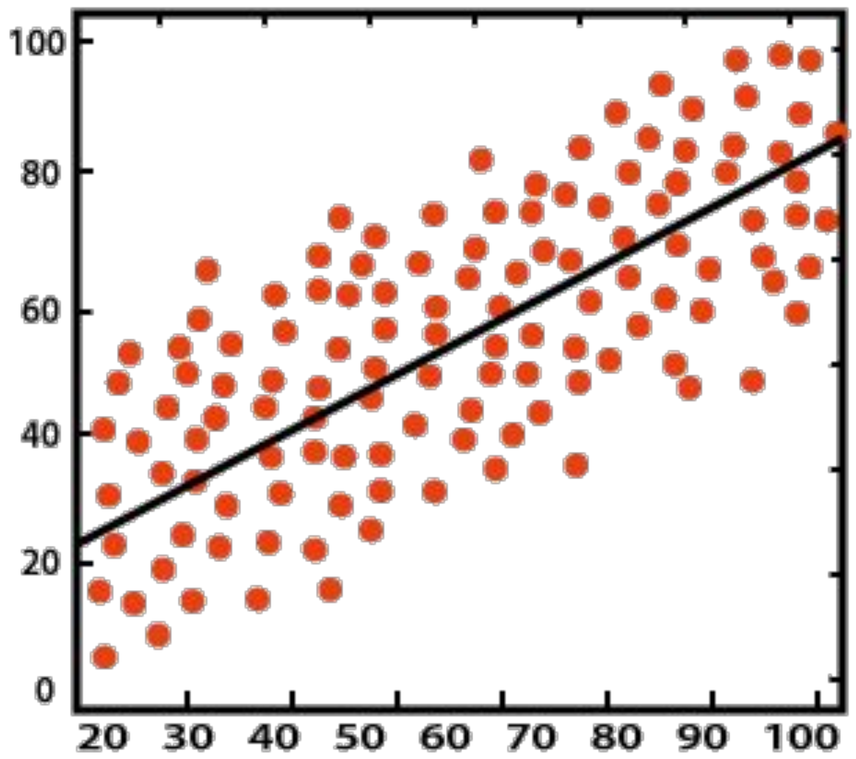


https://www.nature.com/articles/s41576-018-0088-9/figures/1, https://www.nature.com/articles/s41576-018-0088-9/figures/3

Two types of supervised learning



Classification



Regression

Our data

	A	B	C	D	E	F	G	H	I	J	K
1	666942	1	1	1	1	2	1	3	1	1	2
2	667204	7	8	7	6	4	3	8	8	4	4
3	673637	3	1	1	1	2	5	5	1	1	2
4	684955	2	1	1	1	3	1	2	1	1	2
5	688033	1	1	1	1	2	1	1	1	1	2
6	691628	8	6	4	10	10	1	3	5	1	4
7	693702	1	1	1	1	2	1	1	1	1	2
8	704097	1	1	1	1	1	1	2	1	1	2
9	704168	4	6	5	6	7	0	4	9	1	2
10	706426	5	5	5	2	5	10	4	3	1	4
11	709287	6	8	7	8	6	8	8	9	1	4
12	718641	1	1	1	1	5	1	3	1	1	2

- ~700 samples (rows)
  - Training on ~630
  - Testing on ~70
- 11 features (columns)

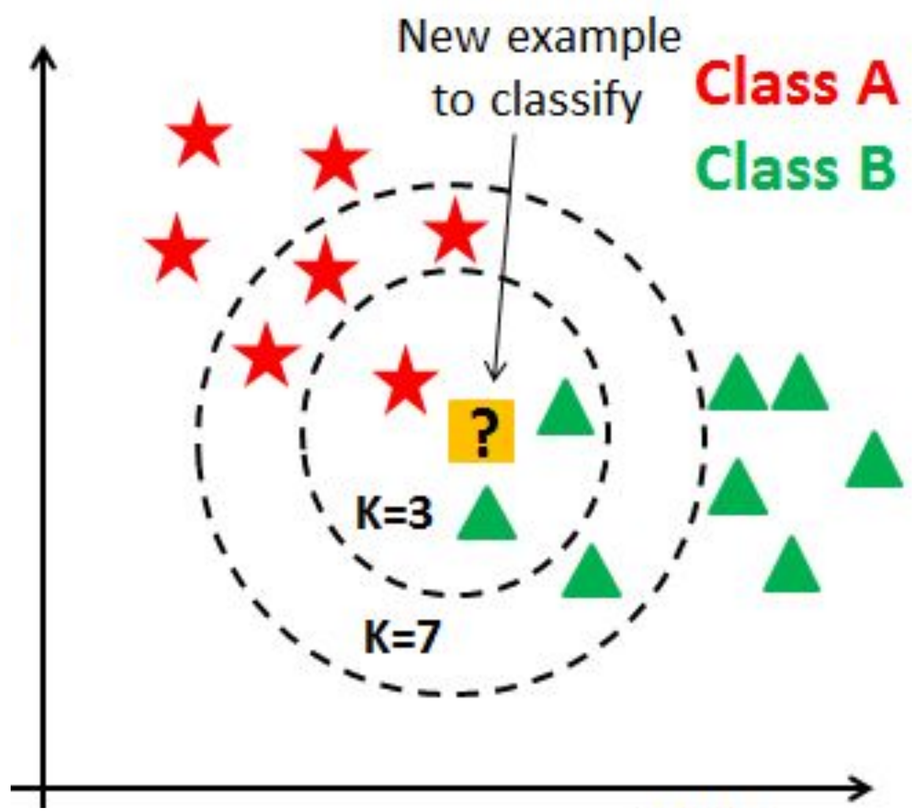
1. Sample code number	id number
2. Clump Thickness	1 - 10
3. Uniformity of Cell Size	1 - 10
4. Uniformity of Cell Shape	1 - 10
5. Marginal Adhesion	1 - 10
6. Single Epithelial Cell Size	1 - 10
7. Bare Nuclei	1 - 10
8. Bland Chromatin	1 - 10
9. Normal Nucleoli	1 - 10
10. Mitoses	1 - 10
11. Class:	(2 for benign, 4 for malignant)

Goal: predict if a cell is benign/malignant based on features 2-10

k-nearest-neighbors algorithm

For each test instance to classify:

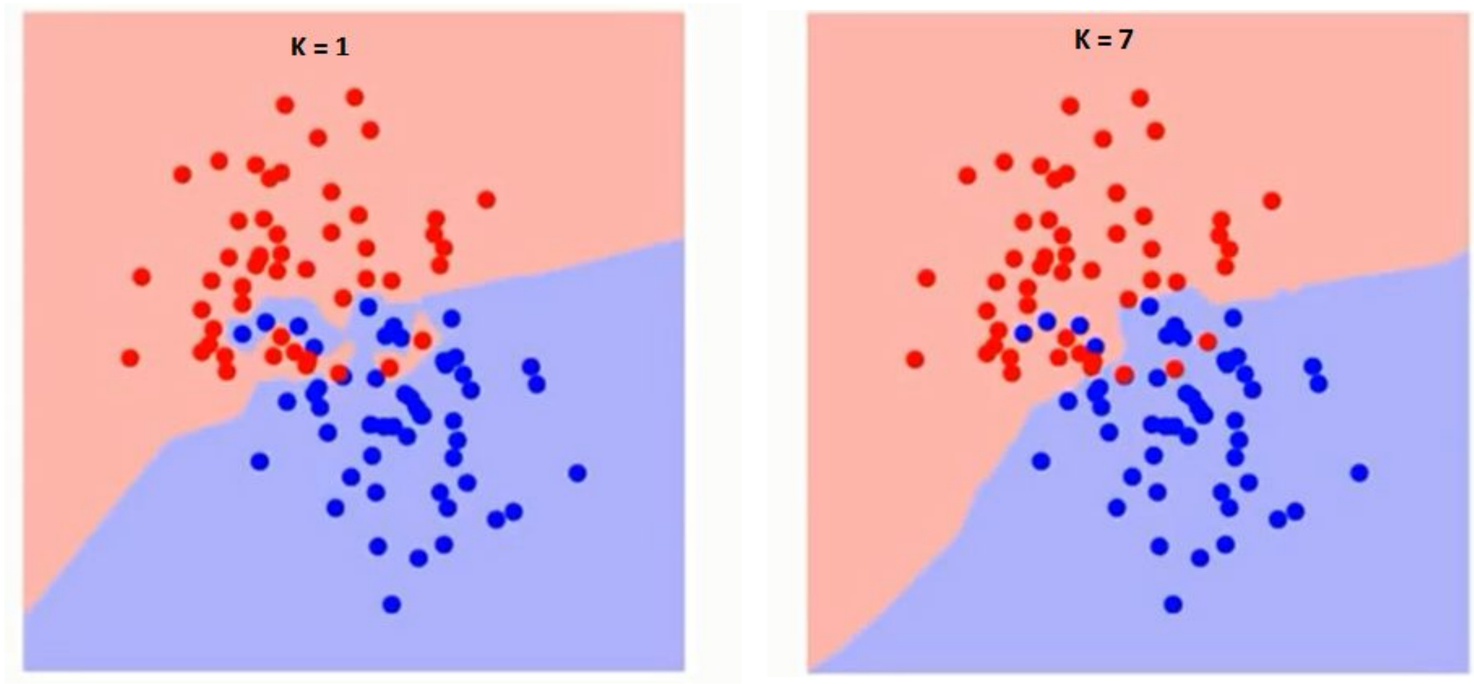
1. Calculate distance between test instance and every train instance
2. Pick the *k* train instances with the smallest distances
3. Of these *k* train instances, see how many are classified as malignant vs. benign
4. Pick whichever class appears more times as your answer



*note: *k* should be odd to avoid ties

k-nearest-neighbors algorithm

For a given *k* value, we can visualize the boundaries of each class



Let’s Get Started!

Slides: [bit.ly/cs106s-knn](bit.ly/cs106s-knn)

Starter code: bit.ly/knn-starter

Files

- **data.js** - loads the dataset
- **index.html** - runs the script **cancer-classify-no-d3.js**
- **cancer-classify-no-d3.js** - script with main logic

Let’s Get Started!

Slides: [bit.ly/cs106s-knn](bit.ly/cs106s-knn)

Starter code: bit.ly/knn-starter

TODOs for you

$$\|x - y\| = \sum_{i=1}^{d} (x_i - y_i)^2$$

1. Implement **calculateDistance()** to calculate L2 distance between points
2. In **kNN()**, find the distance from a test example to all training examples
    - Challenge goal: consider time/space tradeoffs. Keep a topResults object tracking ***only*** the K nearest neighbors’ (1) class and (2) distance from the current test instance. Hint: construct an array of ***objects*** using pushing, popping, and sorting
3. In **kNN()**, use the top *k* examples to “vote” on a result

`sort()`

```
sortableArray = [3, 2, 1]

sortableArray.sort() // OK!

sortOfSortableArray = [23, 8, 15]

sortOfSortableArray.sort() // ????
```

Try it out in the console!

(p.s.: `sort()` is in-place)

sort()

We need to teach `sort()` how to sort.

```
arrayName.sort(function(a, b) {

 return a - b //common: sort by ascending nums

});
```

Return a negative number: a before b

Return a positive number: b before a

Return 0: preserve the original order

Pushing and Popping

`arrayName.pop()` : Remove the last element from an array (in-place!)

`arrayName.push(<thing>)`: Append something to the end of you array! Could be an int, a string, an **object**…… ;) (also in-place!)

- Need to push distance (so we can see if this particular point is in our K Nearest Neighbors!)
- Need to push classification (so we can do the vote!)

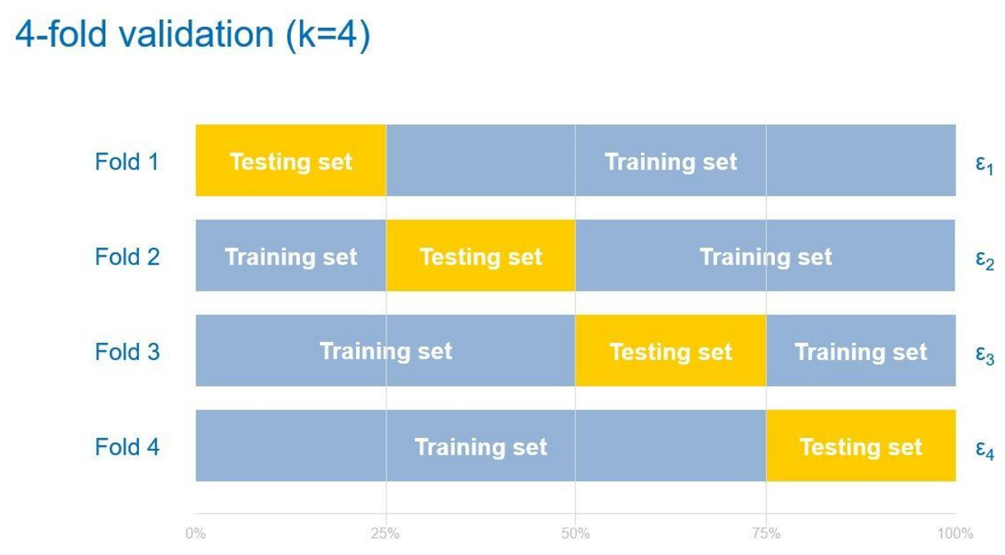
JavaScript Syntax Notes

`var` vs. `let`

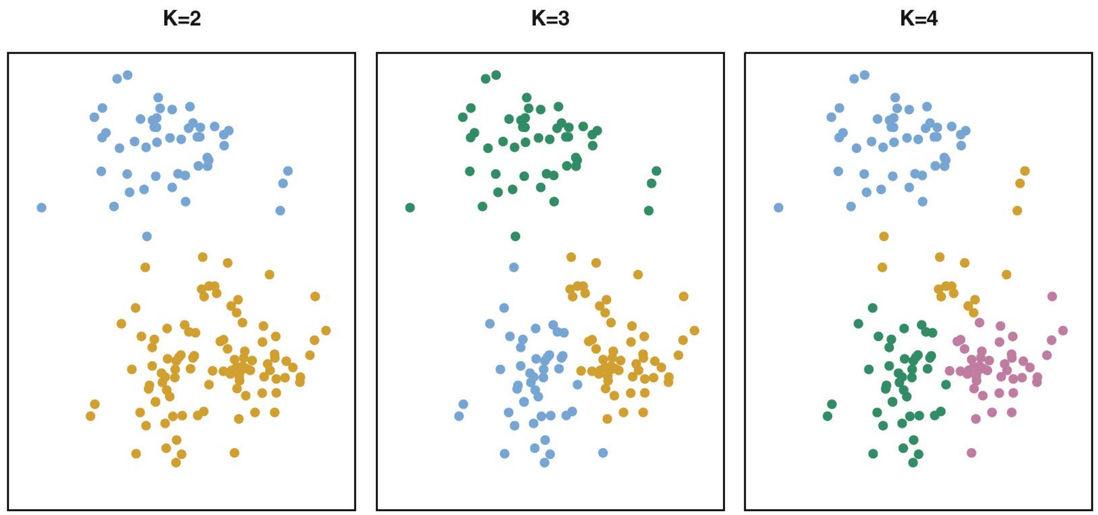
- `let`: variable only available in block where it's defined
- `var`: variable available throughout the function where it's defined

What value of K is best?

- K is a **Hyperparameter**! tl;dr: just gotta play around with it
- But! We only have the one dataset! How can we assess which Ks are good and which ones are bad?
- **Cross-validation**: Redefine what is part of training and what is part of test, then test each K on many training/testing pairs



In contrast to the K-Means algorithm



Algorithm 10.1 *K-Means Clustering*

1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.
2. Iterate until the cluster assignments stop changing:
   (a) For each of the K clusters, compute the cluster *centroid*. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.
   (b) Assign each observation to the cluster whose centroid is closest (where *closest* is defined using Euclidean distance).

Objective Function:

Ensure that clusters are as similar as possible (measured by Euclidean distance in *p*-space)

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}.$$

Reference: An Introduction to Statistical Learning (7e)

More is covered in CS 221, CS 229, Stats 202, Stats 216

Further resources

- Real world connection: “Deep learning-based classification of breast cancer cells using transmembrane receptor dynamics” (Kim et al. 2022) applies deep learning to assess the metastatic potential of breast cancer cells
- More on KNN from CS231N course notes

Checkoff Form

CS 106S